DOCUMENT RESUME

ED 038 165 PS 002 829

AUTHOR Messick, Samuel
TITLE Evaluation of Educational Programs as Research on Educational Process.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE 69
NOTE 11p.; Paper presented at the meeting of the American Psychological Association, Washington, D.C., 1969

EDRS PRICE EDRS Price MF-$0.25 HC-$0.65
DESCRIPTORS *Compensatory Education Programs, Educational Improvement, *Evaluation Methods, *Program Evaluation, Research Methodology, *Research Needs

ABSTRACT

Because of the pressure towards immediate implementation of innovative educational programs, evaluation emphasis has been put on the overall effectiveness of these programs. This type of research yields only yes-no answers about general effectiveness without probing into the process of education itself. More research is needed on the factors that make programs effective; this research can be undertaken without sacrificing action-orientation, by carrying cut the research and the program simultaneously. Evaluative research should include assessment of both possible and intended outcomes, measurements of antecedent conditions and consequences of intervention. If this alteration in research orientation takes place, it could be seen as a shift from the engineering model of evaluation studies (input-output differences relative to cost) to the medical model (concern for specific processes). A major advantage of the medical model is that goals, side effects and program by-products receive increased attention. Finally, this type of evaluative research can serve to advance science as well as social welfare. (MH)

Evaluation of Educational Programs As Research on Educational Process

Samuel Messick
Educational Testing Service

During the past five years, a variety of pressing social problems have been attacked through the medium of innovative and compensatory education programs meant primarily to improve the lot of the culturally disadvantaged and the educationally alienated, with ultimate impact intended not only for students but for families, schools, and communities as well. These programs have been initiated, however, in a political climate marked by fierce competition for scarce economic resources. Given the seriousness of the problems and the scarceness of the resources, it is not surprising that demands should arise from several segments of society that these programs be both timely and effective. As a nation, we simply cannot afford to postpone action pending the results of preliminary research and development efforts that might increase the likelihood of positive effects. Nor can we afford the wasted time and resources of ineffectual programs. In H. G. Wells's[1] aptly prophetic phrasing, "Human history becomes more and more a race between education and catastrophe." Time is of the essence, but rapid intervention is not enough--it must also be effective intervention.

Such pressures toward immediate program implementation have helped to create in some quarters an intellectual--or, more precisely, an anti-intellectual--atmosphere in which research is seen as a frill. In this atmosphere, the watchwords are action and accountability: primary concern is with the initiation and execution of programs and with demonstrating their overall effectiveness. The main question is whether or not the program works. With

---

This paper was presented as part of the Symposium on "Early Learning and Compensatory Education," at the meeting of the American Psychological Association, Washington, D. C., 1969.

time running out, there is little inclination to pause to inquire why it works or even how it works, let alone to ask what aspects of the program work best and for what kinds of students under what kinds of circumstances. Under such conditions, program efforts can easily come to be animated more by a concern for vindication than a concern for verification,[2] with more energy expended in justifying than in judging. But an unnecessarily high price is exacted for this emphasis upon pay-off to the exclusion of process --we do not obtain information about the dynamics of the program or the functioning of its components that could be used to improve the program or to modify it if conditions change.[3]

Fortunately, there is an alternative, and that is to undertake <u>evaluative research</u>. In other words, we should go ahead and do research on program functioning, to acquire information serving the ends of both accountability and program improvement, but within a time frame that will not delay program implementation. This may be accomplished by carrying out the research and the program simultaneously, by including within the administration of the program provision for collecting information relevant to its evaluation and improvement. In some educational programs it is even possible to embed the evaluative research directly into the program itself by capitalizing upon certain kinds of information, such as measures of pupil progress, for both evaluative and instructional purposes. Care must be taken, of course, to ensure that the evaluative research activities do not interfere with or contravene any program activities or intentions. For this reason, heavy reliance should be placed wherever possible upon what Webb and his colleagues have called "unobtrusive measures" of program effects.[4]

If evaluative research is to provide not only a basis for satisfying accountability concerns but also a base of knowledge and understanding that would permit program improvement, extrapolation to other settings and problem areas, and responsiveness to changing conditions, then the form it takes and the kinds of questions it asks must go far beyond the typical engineering model of evaluation studies. The engineering model focuses upon input-output differences relative to cost. This provides information necessary for assessing overall program significance or impact but not sufficient for program revision or development. As Scriven has emphasized, the medical model is a more appropriate paradigm for educational research, and there are several important consequences of this view.[5]

To begin with, there is the recognition that a prescription for treatment and the evaluation of its effectiveness should take into account not only the reported symptoms but other aspects of the organism and its ecology as well. In the present context, this is essentially an affirmation of the need for a systems approach in educational evaluation that would attempt to deal empirically with the interrelatedness of psychological, social, environmental, and educational factors. In the area of human development, we are faced with a particularly complicated system composed of differentiated but overlapping subsystems that embrace family, peer, and community influences as well as school, teacher, and program influences. In such a situation it is possible that compensating trade-offs among variables will occur under different conditions to produce similar effects and that particular outcomes will frequently be multiply determined and sometimes overdetermined. One implication of all this is that evaluative research under such complex circumstances should routinely adopt a multivariate interactional strategy: it should employ

multiple measures in each domain and methods of analysis sensitive to interacting influences. The evaluation question thus becomes elaborated within this framework from a straightforward comparison like "Is this new treatment more effective than the old one?" to something more complicated, like "Do these treatments or treatment components interact with personality and cognitive characteristics of the students or with factors in their educational history or family backgrounds to produce differential effects upon achievement?" Note that such an elaboration is important, for if personality-by-treatment interactions occur or if background factors turn out to moderate treatment effects, then a simple comparison of average gains for different treatment groups will very likely be misleading.[6]

Another derivative of the medical model is a concern for monitoring possible side effects of the treatment. This also follows naturally from the system conception, for if the various elements and subsystems are interdependent, then a change in one part of the system may produce unanticipated and possibly adverse consequences in another part of the system. Because of this possibility, it is not enough to evaluate a program solely in terms of its stated goals, on the basis of how well it achieves its intended objectives. In addition to the intended outcomes, we should also assess a wide range of possible outcomes, for we might unearth in the process some alternatives that ought to be weighed in reaching a final appraisal of program impact. As Henry Dyer has emphasized, "Evaluating the side effects of an educational program may be even more important than evaluating its intended effects."[7]

Another implication of this medical analogy is that feelings and reactions should be assessed periodically throughout the course of the treatment and not just at the beginning and the end. This assessment should include,

as is the custom in medical practice, a monitoring of attitudes toward the treatment itself.

And, of course, underlying the entire metaphor is the notion that whenever possible in the evaluation of educational programs, as in the evaluation of drugs, we should go beyond a simple assessment of the size of effects to an investigation of the processes that produce the effects, for an understanding of these processes will provide a rational basis for changing programs if conditions change and for isolating possible danger zones where potential side effects should be monitored.

In this view, then, evaluative research should focus not only upon the outcomes of education but also upon the process and the context of education, thereby encompassing within its purview several broad areas of measurement concern--input, context, process, and output.[8] In addition, it is probably wise to extend the range of the evaluation process even further in time. As in a medical case study, measures of antecedent conditions should be included, as well as follow-up measures of the consequences both of the treatment and of the termination of treatment. This approach thus emphasizes the importance of comparative longitudinal data in evaluating the effectiveness of educational treatments and stresses the need for analytical procedures that properly take into account those student-process-environment interactions that produce differential results. For these purposes, the scope of measurement must be broad enough both to ensure adequate coverage of potentially interactive variables and to permit the monitoring of possible side effects of the educational programs, particularly in the affective and motivational domains. If possible, measures should be included to assess not only characteristics of the learners but also of their learning environments

(including the home and community as well as the classroom and school), and of the educational processes at all levels (including characteristics of teachers, programs, and classroom dynamics).[9] In this latter connection, it is particularly important to measure characteristics of the program as it is actually carried out, since the program as practiced is sometimes quite different from the program as planned.[10]

Some persons might object at this point that this particular view of evaluative research is overly elaborate and complicated and that with all this emphasis upon process and interacting influences, we are in danger of subverting the main aim of evaluation, which is an appraisal of outcome or product. This latter view has come to be called in some circles the "butter wrapping model" of evaluation.[11] The main thing you want to know, if you have taught someone how to wrap butter, is how many pounds of butter he can wrap. For this purpose, it does not matter that different processes of butter wrapping are employed by different butter wrappers--one may use his right thumb, for example, and another his left forefinger. What matters is how many pounds of butter he can wrap after the training experience. But suppose one very effective butter wrapper has gained his speed and efficiency because of a particular stylistic quirk, namely wetting his thumb with his tongue as he picks up the waxed paper. You may not be very happy with that particular style of butter wrapping even though it produces more wrapped butter than any other. It would seem that at least in some cases, then, it is necessary to take into account factors of style and process in order to evaluate the very desirability of the outcome.

It is thus being suggested here that, wherever possible, evaluative

research should encompass both process and outcome in order to accrue information relevant for both program improvement and administrative decision making. At first glance this may seem like a blurring of the conceptually useful distinction drawn by Scriven between formative evaluation, or evaluation for program improvement, and summative evaluation, or appraisal of the final product, but there really is no conflict involved. The terms "formative" and "summative" refer to different roles of evaluation not to different forms of evaluation, and data from a particular evaluation study can and usually should serve several roles.[12] The research design for the summative evaluation of the Children's Television Workshop series, for example, permits feedback from reactions to shows early in the broadcast year to serve as a guide to the development of shows not scheduled for production until later in the broadcast year.[13] The final product appraised in the summative evaluation is the series of shows actually telecast, but the shape of the latter part of that series may be drastically influenced by evaluations of the earlier part. Thus, it is possible for summative evaluations to serve formative purposes--whether immediately, as in the case of programs developed as a sequence of finished units like Children's Television Workshop, or on a more delayed basis as in periodic program revision--as long as relevant data are collected a sufficient number of times during the course of the program, or under a sufficient number of varying conditions, to permit different program components to be evaluated separately.

Up to this point we have been discussing evaluative research in education primarily as the application of the methods of empirical social science to ascertain the nature and size of the effects of educational treatments.

This covers the first or "descriptive" phase of evaluation, but we should not forget that the evaluation enterprise also includes a second or "judgment" phase, wherein decisions must be made as to whether or not the observed effects attain suitable standards of excellence or acceptability.[14] The methods of social science research are clearly applicable to the descriptive phase, but it is not as generally recognized that they apply to the judgment phase as well. Thus, the concern of evaluative research properly extends to an examination of the criterion or goal scales, for example, and to an appraisal of the weightings and models used for combining performance information judgmentally in administrative decision making. Indeed, the purview of this research even extends to an evaluation of the goals themselves, and particularly to the investigation of the validity of assumptions underlying the goals.[15] Thus, an evaluation of the effectiveness of a program for educating mothers to place their infants on rigid feeding schedules might well involve an investigation of the validity of assumptions relating the expressed goals to underlying values.

As we exercise what Scriven[16] has called "the scientific obligation to evaluation," we should recognize that social science faces not only obligation but opportunity and that evaluative research can serve to advance science as well as the social welfare. As Suchman has so well summarized in his excellent treatise on the topic:

> To some extent evaluative research may offer a bridge between "pure" and "applied" research. Evaluation may be viewed as a field test of the validity of cause-effect hypotheses in basic science....Action programs in any professional field should be based upon the best available scientific knowledge and theory of that field. As such, evaluations of the success or failure of these programs are intimately tied to the proof or disproof of such knowledge. Since such a knowledge

base is the foundation of any action program, the evaluative research worker who approaches his task in the spirit of testing some theoretical proposition rather than a set of administrative practices will in the long run make the most significant contribution to program development.[17]

Footnotes

1. Wells, H. G. The Outline of History (Rev. Ed.) New York: Doubleday, 1956.

2. Chapin, F. S. Experimental Designs in Sociological Research. New York: Harper and Bros., 1947.

3. Cronbach, L. J. Course improvement through evaluation. Teachers College Record, 1963, 64, 672-683.

4. Webb, E. J., Campbell, D. T., Schwartz, R. D., and Sechrest, L. Unobtrusive Measures: Nonreactive Research in the Social Sciences. Chicago: Rand McNally and Co., 1966.

5. Scriven, M. Student values as educational objectives. Proceedings of the 1965 Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1966, pp. 33-49.

6. Messick, S. The criterion problem in the evaluation of instruction: assessing possible, not just intended outcomes. In M. C. Wittrock & D. E. Wiley (Eds.) The Evaluation of Instruction: Issues and Problems. New York: Holt, Rinehart, and Winston, in press.

7. Dyer, H. S. The discovery and development of educational goals. Proceedings of the 1966 Invitational Conference on Testing Problems. Princeton, N.J.; Educational Testing Service, 1967, pp. 12-24; see also Messick, S., in The Evaluation of Instruction: Issues and Problems, in press.

8. Stufflebeam, D. L. Toward a science of educational evaluation. Educational Technology, July 30, 1968, 5-12; Worthen, B. R. Toward a taxonomy of evaluation designs. Educational Technology, August 15, 1968, 3-9.

9. Anderson, Scarvia B., & Doppelt, J. (Chairmen). Untangling the Tangled Web of Education. Research and measurement considerations related to assessing children's development in interaction with school, family, and community influences. Research Memorandum RM-69-6. Princeton, N.J.: Educational Testing Service, 1969; Metfessel, N.S., and Michael, W.B. A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs. Educational and Psychological Measurement, 1967, 27, 931-943.

10. Stake, R. E. The countenance of educational evaluation. Teachers College Record, 1969, 68, 523-540.

11. M. C. Wittrock and D. E. Wiley, (Eds.) The Evaluation of Instruction: Issues and Problems. New York: Holt, Rinehart, and Winston, in press.

12. Scriven, M. The methodology of evaluation. In Perspectives of Curriculum Evaluation. American Educational Research Association Monograph Series on Curriculum Evaluation. Chicago: Rand McNally & Co., 1967, pp. 39-83.

13. Evaluating Sesame Street. A Proposal for Children's Television Workshop. Princeton, N.J.: Educational Testing Service, 1969.

14. Stake, R. E., op.cit.

15. Suchman, E. A. Evaluative Research. New York: Russell Sage Foundation, 1967.

16. Scriven, M. The scientific obligation to evaluation. Paper presented at Symposium on Evaluation of Educational Materials and Processes at the meeting of the American Psychological Association, San Francisco, 1968.

17. Suchman, E. A., op.cit.